
# ORALGEN: Some new functionalities of 2003

# ORALGEN: Some new functionalities of 2003

In the month of February, the Oral Pathogen Sequence (ORALGEN) database server responded to 31,561 requests on the *Streptococcus mutans* database alone. Many of these requests came from search engines that are building their databases. However, in this month there were requests from over 150 academic institutions for information on *Streptococcus mutans*.

The ORALGEN databases are specialized databases funded by the National Institute of Dental and Craniofacial Research (NIDCR) within the National Institutes of Health, Bethesda, Maryland. The scope of the project includes molecular information pertaining to oral pathogens, bacterial and viral. Currently, the databases contain the recently sequenced genomes of *Porphyromonas gingivalis* and *Streptococcus mutans*. *Actinobacillus actinomycetemcomitans* and *Treponema denticola* will be added this year.

## Table of contents

## PredPath - Pathway Predictions

**Description**

PredPath is a biochemical pathway prediction tool that dynamically matches a gene to a pathway in a particular organism for which annotation is available. The placement of a gene within a pathway is based on the presence or absence of Enzyme Commission (EC) numbers in a gene record. The EC number is then cross-referenced in a database that associates EC numbers with functional classes and pathways.

**Features**

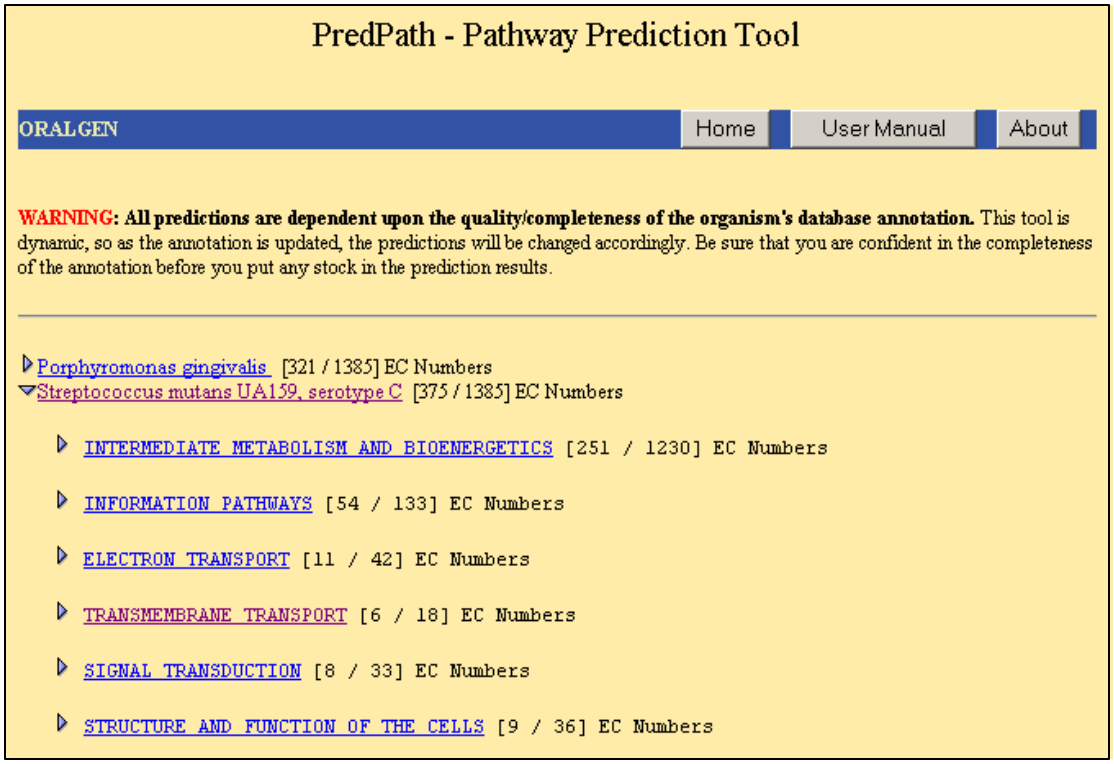
- Browse to discover genes that are part of pathways in an organism and pathways.
- Annotation Aid - seek out missing genes in predicted pathways.
- Search for genes via pathway or find similar genes.

**Access**

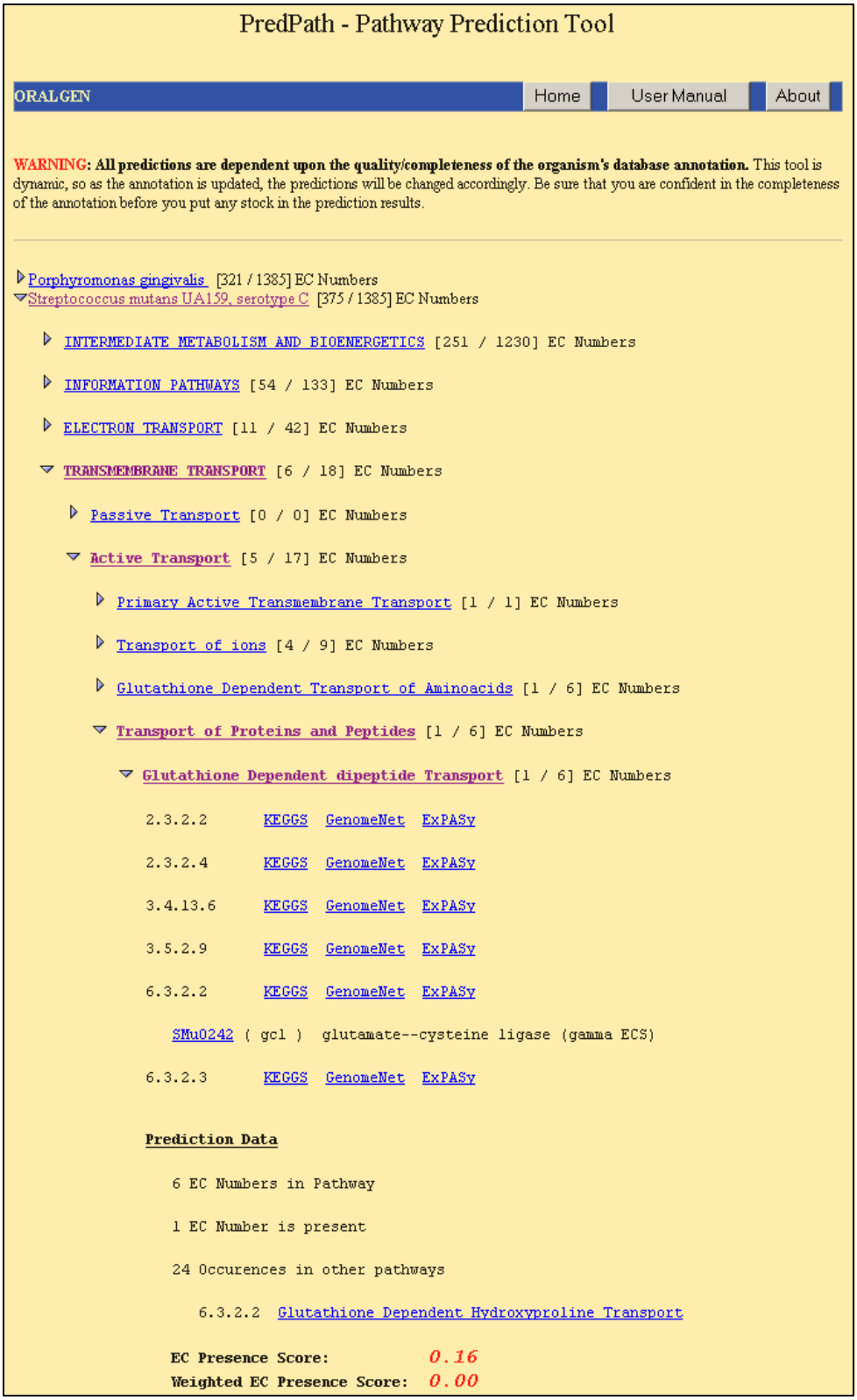
www.oralgen.lanl.gov ➔ **Left frame:** "**Pathway Predictions**" under **"Analysis Tools"**

**Tutorial**

Select "**Pathway Predictions**" tool. In the right frame, select your organism of interest (i.e., *Porphyromonas gingivalis* or *Streptococcus mutans* UA159, serotype C). A listing of the broadest level of functional classes into which the genes in your organism fall will be displayed:




The functional classes become more specific the more you explore the classes. Navigate through the functional classes until you reach your pathway of interest. Use the EC number fraction following the functional class name ([Number of ECs the current organism has / Number of ECs available in our metabolic database in the current category]) as a guide.



PredPath - Pathway Prediction Tool

ORALGEN | Home | User Manual | About

**WARNING: All predictions are dependent upon the quality/completeness of the organism's database annotation.** This tool is dynamic, so as the annotation is updated, the predictions will be changed accordingly. Be sure that you are confident in the completeness of the annotation before you put any stock in the prediction results.

Porphyromonas gingivalis [321 / 1385] EC Numbers
Streptococcus mutans UA159, serotype C [375 / 1385] EC Numbers

- INTERMEDIATE METABOLISM AND BIOENERGETICS [251 / 1230] EC Numbers
- INFORMATION PATHWAYS [54 / 133] EC Numbers
- ELECTRON TRANSPORT [11 / 42] EC Numbers
- TRANSMEMBRANE TRANSPORT [6 / 18] EC Numbers
  - Passive Transport [0 / 0] EC Numbers
  - Active Transport [5 / 17] EC Numbers
    - Primary Active Transmembrane Transport [1 / 1] EC Numbers
    - Transport of ions [4 / 9] EC Numbers
    - Glutathione Dependent Transport of Aminoacids [1 / 6] EC Numbers
    - Transport of Proteins and Peptides [1 / 6] EC Numbers
      - Glutathione Dependent dipeptide Transport [1 / 6] EC Numbers

2.3.2.2 KEGGS GenomeNet ExPASy

2.3.2.4 KEGGS GenomeNet ExPASy

3.4.13.6 KEGGS GenomeNet ExPASy

3.5.2.9 KEGGS GenomeNet ExPASy

6.3.2.2 KEGGS GenomeNet ExPASy

SMu0242 ( gcl ) glutamate--cysteine ligase (gamma ECS)

6.3.2.3 KEGGS GenomeNet ExPASy

**Prediction Data**

6 EC Numbers in Pathway

1 EC Number is present

24 Occurences in other pathways

6.3.2.2 Glutathione Dependent Hydroxyproline Transport

**EC Presence Score:** *0.16*
**Weighted EC Presence Score:** *0.00*

In the example above, the most general functional class chosen was "TRANSMEMBRANE TRANSPORT". The functional classes that follow become more specific (Active Transport ➔ Transport of Proteins and Peptides ➔ Glutathione Dependent dipeptide Transport). Of the 18 possible EC numbers in our metabolic database for the TRANSMEMBRANE TRANSPORT functional class, 6 are referenced in at least one gene record. As you move down the functional class hierarchy, you will come to the most specific functional class 'Glutathione

Dependent dipeptide Transport'. The 6 EC numbers that represent this functional class are located under this class heading. The one that is represented in our genome of interest (*S. mutans*) is near the bottom (6.3.2.2) of this list. The EC numbers are listed in numerical order. Clicking on KEGGS, GenomeNet, or ExPASy will bring you to the respective site where you can get more information on this step in the pathway or view a pathway in its entirety. When you reach a pathway, you will be provided with links to EC information in other metabolic databases and links to genes in the current organism to help you find annotation evidence to support the prediction of the pathway's presence.

The prediction output also tells you how many EC Numbers in the current pathway are used in other pathways. Near the bottom of the display are the "EC Presence score" and the "Weighted EC Presence Score". The fraction provided in the "EC Presence Score" is the number of EC numbers present in the organism for that pathway and the "Weighted EC Presence Score" is the same ratio weighted by two factors: the number of genes with the same EC number and the number of occurrences of an EC number in other pathways. We have adapted the scoring mechanism used by Karp (1999) for our "Weighted EC Presence Score". In this example, you might expect that SMu0242 is not involved in 'Glutathione Dependent dipeptide Transport' because it has an EC presence score of 0.16.

**WARNING: All predictions are dependent upon the quality/completeness of the annotation.** This tool is dynamic, so as the annotation is updated, the predictions will change accordingly.

**Reference**

Karp, P., *et al*. 1999. Integrated pathway-genome databases and their role in drug discovery. Trends Biotechnol. 17(7):275-81.

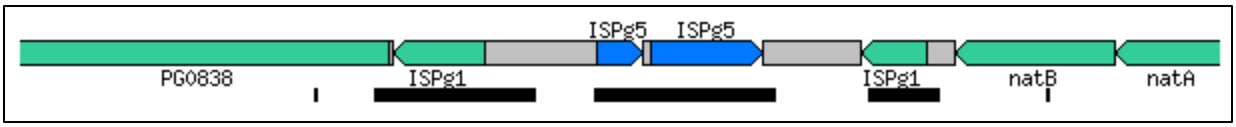
## Repeats

### Description

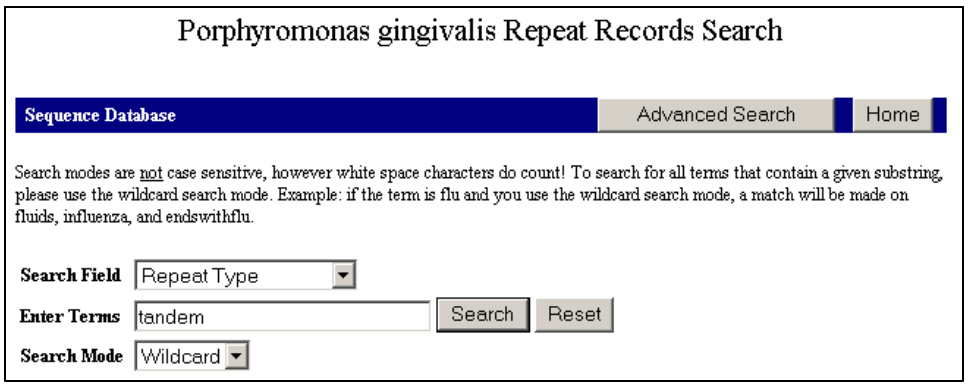
A new table is available for storing information related to repeats. Repeats refer to repetitive segments of DNA. There are various types of repeats (Tandem, Direct, Inverted, IS element). Currently, tandem repeats have been annotated in the ORALGEN database. The tandem repeats were obtained using Tandem Repeat Finder (TRF, http://c3.biomath.mssm.edu/trf.submit.options.html). Direct and inverted repeats were located using REPuter (http://www.genomes.de/).

### Features

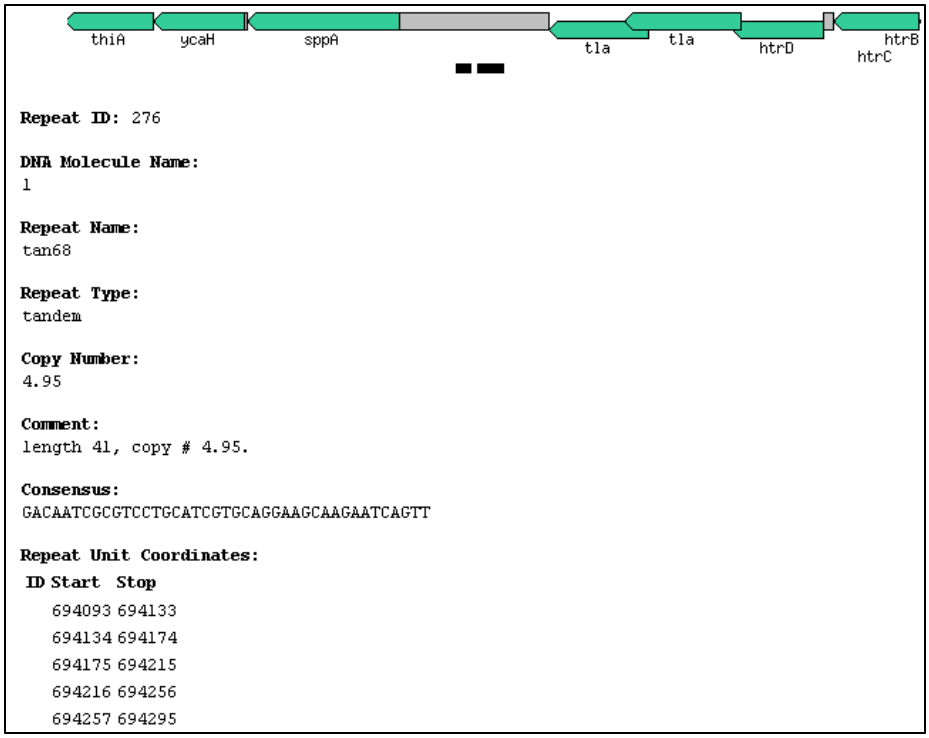
- Organization and placement of repeats can be viewed underneath the gene map within each gene record or on the 'Gene Image Map' as black bars (as seen below). The repeat records are accessible by clicking on the black bars.




- A user-defined search of the repeat table will retrieve and display repeat information according to a **user's query**:




**Result**:




- A predefined database search of the repeat table will retrieve all information for all repeats. This can be accessed in the right frame after an organism is chosen. Here is an excerpt from the repeat table:

**Repeat Table for *Porphyromonas gingivalis***

| Repeat Name | Repeat Type | Copy # | Location | Length | Gene & Intergenic Id(s) | Definitions-Comments |
|---|---|---|---|---|---|---|
| ISPg2 | IS element | 5.00 | 308933,310139<br>927845,929051<br>1430946,1432152<br>1831180,1832386<br>2283295,2284501 | | PG0249 , IGR0210, IGR0211, PG0775 , IGR0660, IGR0661, PG1185 , IGR1017, IGR1018, PG1522 , IGR1299, IGR1300, PG1905 , IGR1635, IGR1636 | ISPg2 (PGIS2) transposase<br>ISPg2 (PGIS2) transposase<br>ISPg2 (PGIS2) transposase<br>ISPg2 (PGIS2) transposase<br>ISPg2 (PGIS2) transposase |
| ISPg3 | IS element | 4.00 | 230914,231983<br>293939,295008<br>852832,853901<br>1339812,1340881 | | PG0178 , IGR0148, IGR0149, PG0236 , IGR0198, IGR0199, PG0718 , IGR0606, IGR0607, PG1115 , IGR0955, IGR0956 | ISPg3 (IS195) transposase<br>ISPg3 (IS195) transposase<br>ISPg3 (IS195) transposase<br>ISPg3 (IS195) transposase |
| tan1 | tandem | 4.33 | 206745,206757 | 3 | PG0167 | hypothetical protein |
| tan2 | tandem | 4.00 | 569898,569909 | 3 | PG0479 | CTP synthetase (UTP ammonia ligase) |
| tan71 | tandem | 2.19 | 1237619,1237708 | 41 | IGR0885 | |
| tan72 | tandem | 2.97 | 1237710,1237831 | 41 | IGR0885 | |

**Access**

www.oralgen.lanl.gov ➔ **Left frame: Analysis** ➔ "**Tandem Repeat Search**"

www.oralgen.lanl.gov ➔ **Left frame:** Your favorite **organism** ➔ Right frame: "**Gene Image Map**"

www.oralgen.lanl.gov ➔ **Left frame:** Your favorite **organism** ➔ Right frame: "**Repeat Search**" under "**User Defined Database Searches"**

www.oralgen.lanl.gov ➔ **Left frame:** Your favorite **organism** ➔ Right frame: "**Repeat Table**" under "**Predefined Database Searches**"

# Sequence Viewer

**Description**

Enables a user to extract flanking sequence 5' and 3' of a gene nucleotide sequence.

**Features**

- Useful for primer design.
- Useful for extracting upstream and downstream sequences that might be used in an analysis of regulatory regions (promoters, activator and repressor binding sites, stem loop structures, etc.).
- Circular-sequence friendly.
- Genomic coordinate display - displays the real coordinates of the gene relative to the genome.
- Output can be customized (number of characters per line, flanking sequence lengths, coordinate display).

**Access**

From within a **gene record** ➔ "**Sequence Viewer**" under "**Gene Nucleotide Sequence"** field

**Tutorial**

Select your organism of interest and search for a gene record using the basic, intermediate or advanced search. Within the gene record, at the field "**Gene Nucleotide Sequence**", there will be a link called "**Sequence Viewer**".

Clicking on this link will bring you to a page that looks like this:

*Porphyromonas gingivalis*

**Gene Flanking Sequence Viewer**

This view is based on a particular gene which is displayed with some associated information above the sequence. You may add on "Flanking sequence" either upstream (Start Flank Size) or downstream (Stop Flank Size). You may specify the number of sequence characters per line and choose whether or not you want coordinates displayed to the right, left, or both left and right of the sequence. The sequence displayed is the sense strand relative to the gene the viewer is displaying. In reading the coordinates, note that there is no zero coordinate and that the genomic sequence is assumed to be circular.

Start Flank Size: 50
Stop Flank Size: 50
Characters per line: 50
Show Coordinates on: Left: ☑ Right: ☑

Submit Reset

Gene ID: PG0198.1
Strand: Reverse Compliment
Gene Start: 254843
Gene Stop: 254535
Gene Length: 309

Pick Primers

```
254893  CGTAAAGAGAGATTGAAAAAATACAAGAATAAATAAAACAGCGACAAGAA  254844
254843  ATGAGTAGTAAAGCATGTAGATGCCCCGTATGTGGAGTAGCAACTGTCGG  254794
254793  AAAGCCCAAAAGAAGCTTTGGCAGTAAGCTAGCGCGCAATACATTGAAAA  254744
254743  AAGGAAAAGGAGCGGCTATAGGGGCTGCTATAGGAAGTGTAATTCCGGGT  254694
254693  CTTGGGACAATCACGGGAGGAATTGCAGGTGCAGCAGTAGAATCTTTGAT  254644
254643  TGGAGAAAAAGCAAATGAAAATATCGACAACTTTGCTGACAATTTTTTCC  254594
254593  AGGACACTGAGTTTGAATTTCAATGCAAAAGTTGTTCTCATAAGTGGACA  254544
254543  AGAAAATACTAGGATAGGCAATCCTTTCCTTTATATAAAGAGCTGATATA  254494
254493  TCCTTTTCT                                          254485
```

Enter numbers into the "**Start Flank Size**" and "**Stop Flank Size**" windows to reflect how many nucleotides you wish to view up- and downstream of your gene. In this case, the number 50 was chosen for both. Also, enter the number of nucleotides you wish to view per line ("**Characters per line**"). In this case, the number 50 was entered. You may also choose to "**Show Coordinates on**" the left, right, both or neither sides of the resulting sequence.

After clicking on "**Submit**", the gene nucleotide sequence flanked by 50 nucleotides at each end will be presented. The coordinates displayed represent those of the gene and flanking sequence relative to the genome. You can see there are 50 nucleotides per line and coordinates are on each side (left and right).

## **Primer3** - Primer design

**Description**

Primer3 is an algorithm used for the design of primers or oligonucleotides (i.e., for sequencing, PCR, hybridization, etc).

**Features**

- Integrated with sequence viewer – click "Pick Primers".
- Automatic entry of sequence with information on definition line including relative gene coordinates and length.
- Automatic entry of the relative start coordinate of your target sequence.
- Dynamic link to MIT's current Primer3 application.

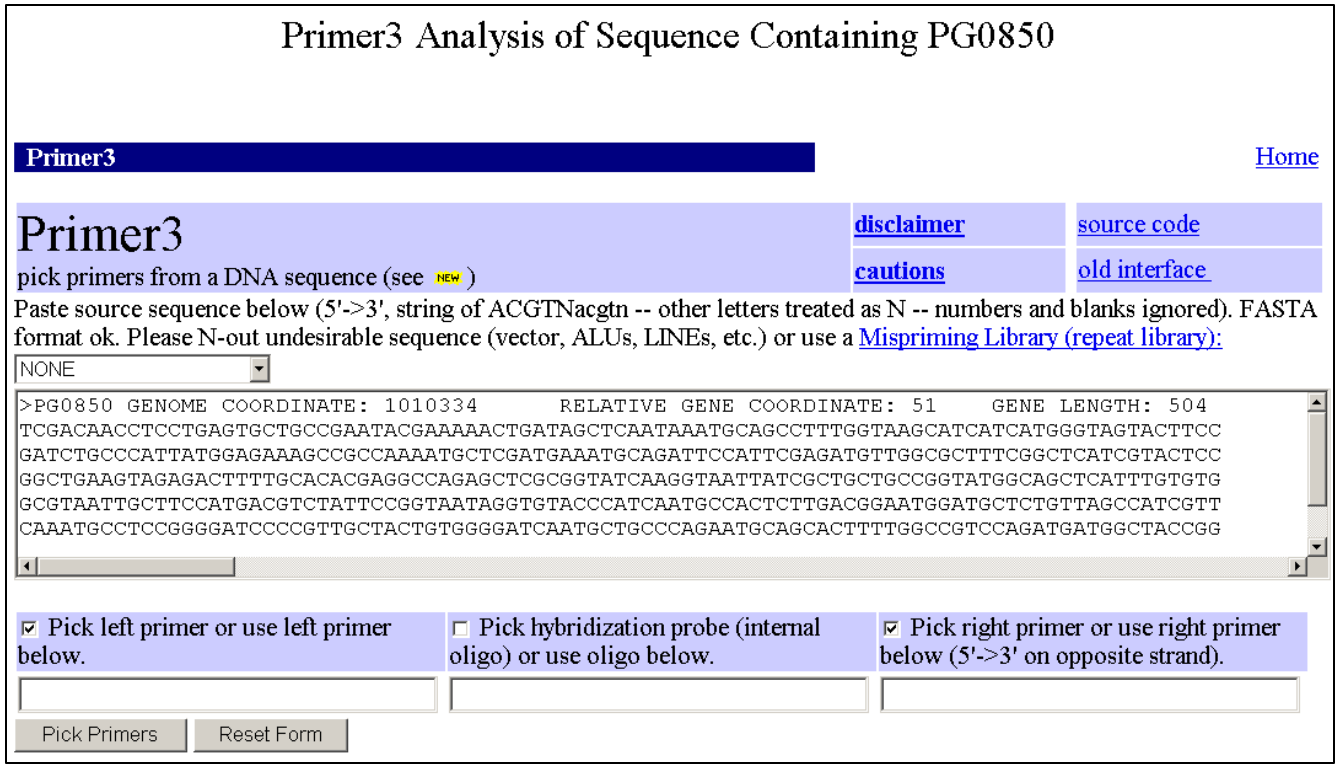
**Access**

From **Sequence Viewer** ➔ Click "**Pick Primers**"

**Tutorial**

After selecting your organism and gene of interest, click on "**Sequence Viewer**" at the "**Gene Nucleotide Sequence**" field. This will bring you to the Sequence Viewer. If you wish to just use your gene nucleotide sequence as the template for primer selection, click on "**Pick Primers**". However, if you wish to include flanking sequence in your template for primer selection, choose the size of the flanking sequence and click "**Submit**". Then click "**Pick Primers**".

Clicking on "**Pick Primers**" will bring you the Primer3 interface and looks like this:




For general primer selection, choose among the three checkboxes labeled "**Pick left primer**…", "**Pick hybridization probe**…", and "**Pick right primer**…". For more selective primer design, you would enter parameters (template targeting, conditions, weighting of primers and primer pairs) in those specific sections.

**Reference**

The development of Primer3 and the Primer3 web site was funded by the Howard Hughes Medical Institute and the National Institutes of Health, National Human Genome Research Institute, under grants R01-HG00257 (to David C. Page) and P50-HG00098 (to Eric S. Lander).

## FTP Site

**Description**

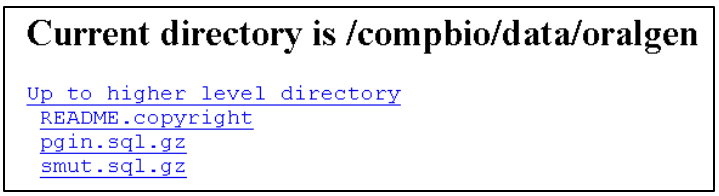
The FTP site is a permanent site that allows downloads of software, MySQL database data dumps and PERL modules. The site is updated regularly and the URL is stable.

**Contents**

- Data – currently available
  - MySQL Database Dumps
- Software – soon to be available

**Access**

**Left frame**: **FTP downloads** (ftp://bpublic.lanl.gov/compbio/data/oralgen/)



## Pre-run PSI-BLAST

**Description**

PSI-BLAST is pre-run on all gene sequences in our database. The top 10 results are displayed when the e-value is below 1e-5. The entire PSI-BLAST result can also be displayed. This feature is useful because it saves time by avoiding a manual PSI-BLAST, bypassing the PSI-BLAST waiting period. The field is updated monthly using the nr database at GenBank.

**Features**

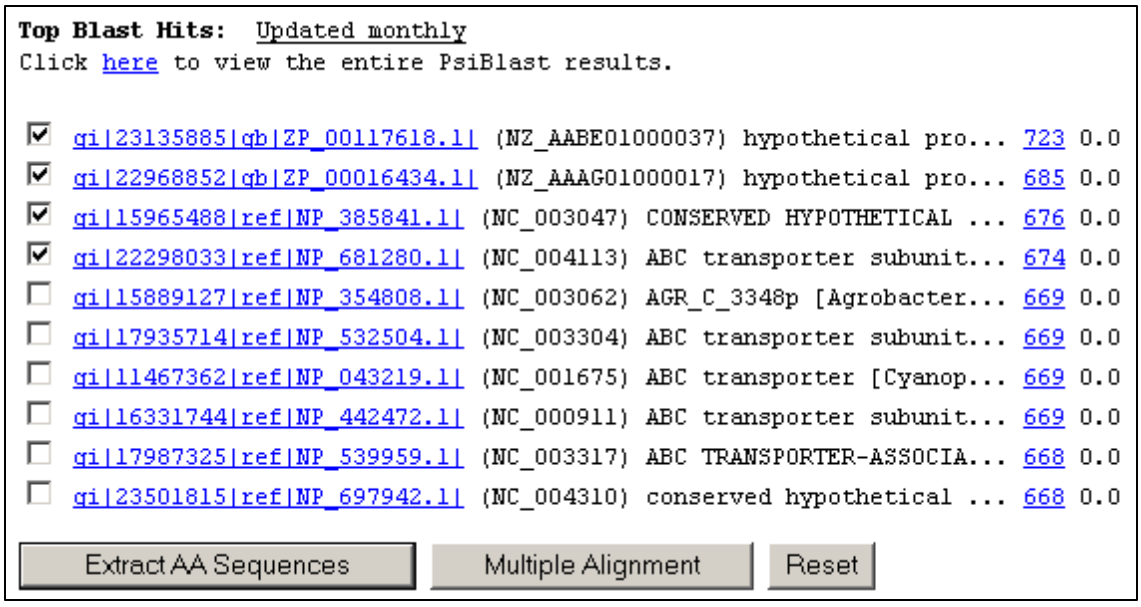
- Either one or multiple PSI-BLAST hits can be selected and a multiple sequence alignment using ClustalW can be performed. The alignment includes the query sequence.
- Amino acid sequences can also be retrieved for the selected BLAST hits.

**Access**

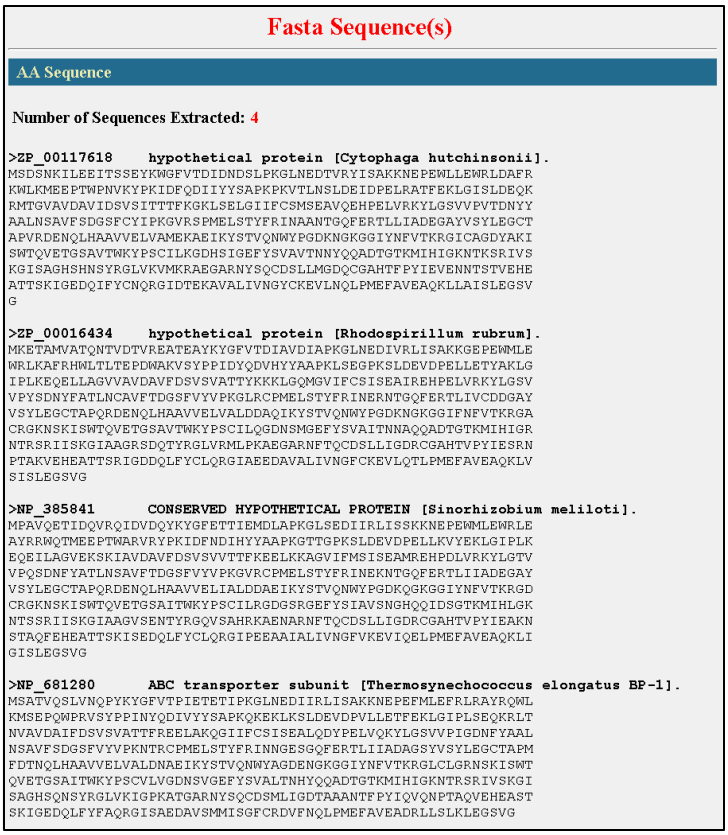
**Left frame**: Your favorite **organisms** ➔ **Search** (Basic, Intermediate, Advanced) ➔ "**Top Blast Hits**" field within a gene record

**Tutorial**

After retrieving a gene record of interest, go to the area of the page listing the "**Top Blast Hits**" field. Under this heading will appear up to 10 BLAST hits to the gene of interest from the nr database. It would look like this:



Clicking on "here" will display the entire PSI-BLAST result. Clicking on any of the hyper-linked records will bring you to that record in GenBank. The button "**Extract AA Sequences**" will return the FASTA-formatted amino acid sequence of records selected, as illustrated here:



**Fasta Sequence(s)**

**AA Sequence**

**Number of Sequences Extracted: 4**

```
>ZP_00117618    hypothetical protein [Cytophaga hutchinsonii].
MSDSNKILEEITSSEYKWGFVTDIDNDSLPKGLNEDTVRYISAKKNEPEWLLEWRLDAFR
KWLKMEEPTWPNVKYPKIDFQDIIYYSAPKPKVTLNSLDEIDPELRATFEKLGISLDEQK
RMTGVAVDAVIDSVSITTTFKGKLSELGIIFCSMSEAVQEHPELVRKYLGSVVPVTDNYY
AALNSAVFSDGSFCYIPKGVRSPMELSTYFRINAANTGQFERTLLIADEGAYVSYLEGCT
APVRDENQLHAAVVELVAMEKAEIKYSTVQNWYPGDKNGKGGIYNFVTKRGICAGDYAKI
SWTQVETGSAVTWKYPSCILKGDHSIGEFYSVAVTNNYQQADTGTKMIHIGKNTKSRIVS
KGISAGHSHNSYRGLVKVMKRAEGARNYSQCDSLLMGDQCGAHTFPYIEVENNTSTVEHE
ATTSKIGEDQIFYCNQRGIDTEKAVALIVNGYCKEVLNQLPMEFAVEAQKLLAISLEGSV
G

>ZP_00016434    hypothetical protein [Rhodospirillum rubrum].
MKETAMVATQNTVDTVREATEAYKYGFVTDIAVDIAPKGLNEDIVRLISAKKGEPEWMLE
WRLKAFRHWLTLTEPDWAKVSYPPIDYQDVHYYAAPKLSEGPKSLDEVDPELLETYAKLG
IPLKEQELLAGVVAVDAVFDSVSVATTYKKKLGQMGVIFCSISEAIREHPELVRKYLGSV
VPYSDNYFATLNCAVFTDGSFVYVPKGLRCPMELSTYFRINERNTGQFERTLIVCDDGAY
VSYLEGCTAPQRDENQLHAAVVELVALDDAQIKYSTVQNWYPGDKNGKGGIFNFVTKRGA
CRGKNSKISWTQVETGSAVTWKYPSCILQGDNSMGEFYSVAITNNAQQADTGTKMIHIGR
NTRSRIISKGIAAGRSDQTYRGLVRMLPKAEGARNFTQCDSLLIGDRCGAHTVPYIESRN
PTAKVEHEATTSRIGDDQLFYCLQRGIAEEDAVALIVNGFCKEVLQTLPMEFAVEAQKLV
SISLEGSVG

>NP_385841      CONSERVED HYPOTHETICAL PROTEIN [Sinorhizobium meliloti].
MPAVQETIDQVRQIDVDQYKYGFETTIEMDLAPKGLSEDIIRLISSKKNEPEWMLEWRLE
AYRRWQTMEEPTWARVRYPKIDFNDIHYYAAPKGTTGPKSLDEVDPELLKVYEKLGIPLK
EQEILAGVEKSKIAVDAVFDSVSVVTTFKEELKKAGVIFMSISEAMREHPDLVRKYLGTV
VPQSDNFYATLNSAVFTDGSFVYVPKGVRCPMELSTYFRINEKNTGQFERTLIIADEGAY
VSYLEGCTAPQRDENQLHAAVVELIALDDAEIKYSTVQNWYPGDKQGKGGIYNFVTKRGD
CRGKNSKISWTQVETGSAITWKYPSCILRGDGSRGEFYSIAVSNGHQQIDSGTKMIHLGK
NTSSRIISKGIAAGVSENTYRGQVSAHRKAENARNFTQCDSLLIGDRCGAHTVPYIEAKN
STAQFEHEATTSKISEDQLFYCLQRGIPEEAAIALIVNGFVKEVIQELPMEFAVEAQKLI
GISLEGSVG

>NP_681280      ABC transporter subunit [Thermosynechococcus elongatus BP-1].
MSATVQSLVNQPYKYGFVTPIETETIPKGLNEDIIRLISAKKNEPEFMLEFRLRAYRQWL
KMSEPQWPRVSYPPINYQDIVYYSAPKQKEKLKSLDEVDPVLLETFEKLGIPLSEQKRLT
NVAVDAIFDSVSVATTFREELAKQGIIFCSISEALQDYPELVQKYLGSVVPIGDNFYAAL
NSAVFSDGSFVYVPKNTRCPMELSTYFRINNGESGQFERTLIIADAGSYVSYLEGCTAPM
FDTNQLHAAVVELVALDNAEIKYSTVQNWYAGDENGKGGIYNFVTKRGLCLGRNSKISWT
QVETGSAITWKYPSCVLVGDNSVGEFYSVALTNHYQQADTGTKMIHIGKNTRSRIVSKGI
SAGHSQNSYRGLVKIGPKATGARNYSQCDSMLIGDTAAANTFPYIQVQNPTAQVEHEAST
SKIGEDQLFYFAQRGISAEDAVSMMISGFCRDVFNQLPMEFAVEADRLLSLKLEGSVG
```

If you choose to do a multiple sequence alignment, select the BLAST hits of interest and click "**Multiple Alignment**". This will bring you to the ClustalW site (shown below).

**Multiple Sequence Alignment (ClustalW)**

**AA Sequence Alignment**

ch.EMBnet.org

**ClustalW**

**Valid format for input is: FASTA(Pearson)**
**max number of sequences = 30**
**max total length of sequences = 10000**

Help page

| | | | |
|---|---|---|---|
| Scoring matrix : | Blosum | | |
| Opening gap penalty : | 10 | Extending gap penalty : | 0.05 |
| End gap penalty : | 10 | Separation gap penalty : | 0.05 |
| Output format : | Clustal | Output order : | Input |

Input sequences:
(see above for valid formats)

```
>>>PG0232
MQENNNILDEVTGSEYKYGFVTDIETETIGRGLSEDTVRLISAKKEEPEW
>ZP_00117618
MSDSNKILEEITSSEYKWGFVTDIDNDSLPKGLNEDTVRYISAKKNEPEW
KWLKMEEPTWPNVKYPKIDFQDIIYYSAPKPKVTLNSLDEIDPELRATFE
RMTGVAVDAVIDSVSITTTFKGKLSELGIIFCSMSEAVQEHPELVRKYLG
AALNSAVFSDGSFCYIPKGVRSPMELSTYFRINAANTGQFERTLLIADEG
APVRDENQLHAAVVELVAMEKAEIKYSTVQNWYPGDKNGKGGIYNFVTKR
SWTQVETGSAVTWKYPSCILKGDHSIGEFYSVAVTNNYQQADTGTKMIHI
KGISAGHSHNSYRGLVKVMKRAEGARNYSQCDSLLMGDQCGAHTFPYIEV
```

Run ClustalW

Enter your parameters and click "**Run ClustalW**". A multiple sequence alignment will then be performed using ClustalW. Choose "**clustalw (aln)**" from the ClustalW results page to view the alignment:

```
CLUSTAL W (1.74) multiple sequence alignment


>>PG0232          --------MQENNNILDEVTGSEYKYGFVTDIETETIGRGLSEDTVRLISAKKEEPEWLL
ZP_00117618       --------MSDSNKILEEITSSEYKWGFVTDIDNDSLPKGLNEDTVRYISAKKNEPEWLL
ZP_00016434       MKETAMVATQNTVDTVREA-TEAYKYGFVTDIAVDIAPKGLNEDIVRLISAKKGEPEWML
NP_385841         -----MPAVQETIDQVRQIDVDQYKYGFETTIEMDLAPKGLSEDIIRLISSKKNEPEWML
NP_681280         -----------MSATVQSLVNQPYKYGFVTPIETETIPKGLNEDIIRLISAKKNEPEFML
                                  : .   . **:** * *  :   :**.** :* **:** ***::*


>>PG0232          EFRLNAYRHWLSMKEPDWAHLNIPPIDYQDIIYYAAPKKKKGPKSLDEVDPELLKTFDKL
ZP_00117618       EWRLDAFRKWLKMEEPTWPNVKYPKIDFQDIIYYSAPKPKVTLNSLDEIDPELRATFEKL
ZP_00016434       EWRLKAFRHWLTLTEPDWAKVSYPPIDYQDVHYYAAPKLSEGPKSLDEVDPELLETYAKL
NP_385841         EWRLEAYRRWQTMEEPTWARVRYPKIDFNDIHYYAAPKGTTGPKSLDEVDPELLKVYEKL
NP_681280         EFRLRAYRQWLKMSEPQWPRVSYPPINYQDIVYYSAPKQKEKLKSLDEVDPVLLETFEKL
                  *:** *:*:* .: ** *..:  * *:::*: **:*** .   :****:** *  .: **


>>PG0232          GIPLEEQKILSGM-----AVDAVMDSVSVKTTFKEKLAEKGIIFCSFSEAVKDFPDLVKK
ZP_00117618       GISLDEQKRMTGV-----AVDAVIDSVSITTTFKGKLSELGIIFCSMSEAVQEHPELVRK
ZP_00016434       GIPLKEQELLAGV----VAVDAVFDSVSVATTYKKKLGQMGVIFCSISEAIREHPELVRK
NP_385841         GIPLKEQEILAGVEKSKIAVDAVFDSVSVVTTFKEELKKAGVIFMSISEAMREHPDLVRK
NP_681280         GIPLSEQKRLTNV-----AVDAIFDSVSVATTFREELAKQGIIFCSISEALQDYPELVQK
                  **.*.**: ::.:     ****::****: **:: :* : *:** *:***:::.*:**:*


>>PG0232          YLGTVVSSKDNFFAALNSAVFSDGSFVYIPKGVRCPMELSTYFRINAANTGQFERTLIVA
ZP_00117618       YLGSVVPVTDNYYAALNSAVFSDGSFCYIPKGVRSPMELSTYFRINAANTGQFERTLLIA
ZP_00016434       YLGSVVPYSDNYFATLNCAVFTDGSFVYVPKGLRCPMELSTYFRINERNTGQFERTLIVC
NP_385841         YLGTVVPQSDNFYATLNSAVFTDGSFVYVPKGVRCPMELSTYFRINEKNTGQFERTLIIA
NP_681280         YLGSVVPIGDNFYAALNSAVFSDGSFVYVPKNTRCPMELSTYFRINNGESGQFERTLIIA
                  ***:**.  **::*:**.***:**** *:**. *.********** ::*******:::.


>>PG0232          DEDSYVSYLEGCTAPQRDENQLHAAIVEIIAETNAEVKYSTVQNWYPGDKEGKGGIYNFV
ZP_00117618       DEGAYVSYLEGCTAPVRDENQLHAAVVELVAMEKAEIKYSTVQNWYPGDKNGKGGIYNFV
ZP_00016434       DDGAYVSYLEGCTAPQRDENQLHAAVVELVALDDAQIKYSTVQNWYPGDKNGKGGIFNFV
NP_385841         DEGAYVSYLEGCTAPQRDENQLHAAVVELIALDDAEIKYSTVQNWYPGDKQGKGGIYNFV
NP_681280         DAGSYVSYLEGCTAPMFDTNQLHAAVVELVALDNAEIKYSTVQNWYAGDENGKGGIYNFV
                  * .:***********  * ******:**::*  .*::*********.**::*****:***


>>PG0232          TKRGVCKGDNSKISWTQVETGSAITWKYPSCVLRGDNSIAEFYSVAVTNNFQQADTGTKM
ZP_00117618       TKRGICAGDYAKISWTQVETGSAVTWKYPSCILKGDHSIGEFYSVAVTNNYQQADTGTKM
ZP_00016434       TKRGACRGKNSKISWTQVETGSAVTWKYPSCILQGDNSMGEFYSVAITNNAQQADTGTKM
NP_385841         TKRGDCRGKNSKISWTQVETGSAITWKYPSCILRGDGSRGEFYSIAVSNGHQQIDSGTKM
NP_681280         TKRGLCLGRNSKISWTQVETGSAITWKYPSCVLVGDNSVGEFYSVALTNHYQQADTGTKM
                  **** * *  :************:********:* ** * .****:*::*  ** *:****


>>PG0232          IHLGKNTRSRIVSKGISAGSSQNSYRGLVKISKNAVNARNHSQCDSLLLSDHCGAHTYPY
ZP_00117618       IHIGKNTKSRIVSKGISAGHSHNSYRGLVKVMKRAEGARNYSQCDSLLMGDQCGAHTFPY
ZP_00016434       IHIGRNTRSRIISKGIAAGRSDQTYRGLVRMLPKAEGARNFTQCDSLLIGDRCGAHTVPY
NP_385841         IHLGKNTSSRIISKGIAAGVSENTYRGQVSAHRKAENARNFTQCDSLLIGDRCGAHTVPY
NP_681280         IHIGKNTRSRIVSKGISAGHSQNSYRGLVKIGPKATGARNYSQCDSMLIGDTAAANTFPY
                  **:*:** ***:****:** *.::*** *    .* .***.:****:*:.* ..*:* **


>>PG0232          ADVQNDTAIIEHEATTSKISEEQIFYCNQRGIGTEEAVGLIVNGYAKEVMNKLLMEFAVE
ZP_00117618       IEVENNTSTVEHEATTSKIGEDQIFYCNQRGIDTEKAVALIVNGYCKEVLNQLPMEFAVE
ZP_00016434       IESRNPTAKVEHEATTSRIGDDQLFYCLQRGIAEEDAVALIVNGFCKEVLQTLPMEFAVE
NP_385841         IEAKNSTAQFEHEATTSKISEDQLFYCLQRGIPEEAAIALIVNGFVKEVIQELPMEFAVE
NP_681280         IQVQNPTAQVEHEASTSKIGEDQLFYFAQRGISAEDAVSMMISGFCRDVFNQLPMEFAVE
                  : .* *: .****:**:*.::*:**  ****  * *:.:::.*: ::*:: * ******


>>PG0232          AQKLLSISLEGSVG
ZP_00117618       AQKLLAISLEGSVG
ZP_00016434       AQKLVSISLEGSVG
NP_385841         AQKLIGISLEGSVG
NP_681280         ADRLLSLKLEGSVG
                  *::*:.:.******
```

## Acknowledgements

ORALGEN, STDGEN, and CBNP teams:
**Dennis F. Mangan, Ph.D.** - Branch Chief and Project Officer DER/NIDCR/NIH
**Gerry Myers & Thomas Brettin** - Principal Investigators

**Cathy Cleland** – Programmer/Annotator/Bioinformatician
**Jonathan Eng** - Student Annotator
**Robert Leach** - Programmer
**Monica Misra** - Student Annotator
**Kim Prichard** - Annotator/Programmer
**Jian Song** - Bioinformatician/Programmer
**Chris Stubben** - Graduate Student, New Mexico State U, Database Administrator
**Nina Thayer** - Annotator/Programmer
**Gary Xie** - Postdoctoral Fellow in Bioinformatics